Közeli rokonunk, az autó

Siklósi Borbála[1], Novák Attila[1,2]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar,
[2] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter utca 50/a
e-mail:{siklosi.borbala,novak.attila}@itk.ppke.hu

Kivonat Számos nyelvtechnológiai probléma megoldására sikerrel alkalmazhatóak a különböző statisztikai módszerek. A fogalmak szemantikai reprezentációja esetén azonban még mindig sokszor az előre, kézzel létrehozott lexikai erőforrásokra támaszkodunk. Cikkünkben azt mutatjuk be, hogy a disztribúciós szemantika modelljét alkalmazva, szöveges korpuszból hogyan nyerhetők ki releváns fogalmi csoportok automatikus módszerekkel. Az algoritmust több különböző doménből származó magyar nyelvű részkorpuszra alkalmaztuk. Az eredmények bebizonyították, hogy a módszer alkalmas az általános értelemben kapcsolódó fogalmak csoportosítása mellett a lazább és asszociációs relációk felismerésére is, illetve jól kezeli a különböző domének közötti hangsúlybeli különbségeket is.

1. Bevezetés

A *big data* és a gyakorlatilag végtelen kapacitású számítógépek korában a nyelvtechnológiai alkalmazások is egyre inkább a statisztikai módszerekhez nyúlnak. Ennek ellenére, a világismeret és a szemantikai relációk ábrázolása általában még mindig kézzel készült lexikai erőforrások (pl. WordNet[4,12]) és doménspecifikus tezauruszok segítségével történik.

Zhang [19] tanulmányában kimutatta, hogy gyakran a fogalmak ilyen mesterséges rendszerezése és a kognitív emberi tudásreprezentáció között jelentős különbség van. Ezért a tudásábrázolás létrehozásakor érdemes lehet a szöveges adatokból kiindulni, ahelyett, hogy ezeket próbálnánk egy előredefiniált fogalomrendszerhez való illeszkedésre kényszeríteni. Ráadásul, kisebb nyelvek esetén, mint a magyar, jóval kevesebb és kisebb lefedettségű lexikai erőforrás áll rendelkezésre, mint a nagyobb nyelveknél [9].

Jelen cikkünkben olyan kísérletekről számolunk be, amelyek azonos fogalmak különböző jellegű szövegekben való használatát vizsgálják. Olyan statisztikai módszereket alkalmaztunk, melyek magyar szövegekből kinyert fogalmakat és kifejezéseket szemantikai csoportokba sorolnak az adott doménen belül való disztribúciós viselkedésük alapján.

A WordNet jellegű erőforrásokban a fogalmakat szinonimahalmazok (synsetek) reprezentálják, melyek az azonos jelentésű szavakat foglalják össze. Ezek között a halmazok között állhatnak fenn explicit relációk. Disztribúciós modellek használatával ezeket a relációkat nem tudjuk automatikusan azonosítani, és

az automatikusan létrejövő szóhalmazok is tartalmaznak oda nem illeszkedő szóalakokat. Ugyanakkor az eredmények azt mutatják, hogy a modell valóban összetartozó kifejezéseket ismer fel, csupán az összetartozás szemantikai típusa tér el akár egy csoporton belül. A kinyert hasonlóságok jellege paradigmatikus, azaz a hasonló kifejezések egymással felcserélhetőek, de ez épp úgy igaz lehet szinonimákra, hipernimákra, hiponimákra és akár antonimákra is. Ezek megkülönböztetésére nem alkalmas a módszer. Ennek ellenére a létrejövő fogalmi csoportok relevánsak. Sőt, azok az asszociációs relációk, melyek a modell alkalmazásával felismerhetők, gyakran hiányoznak a klasszikus ontológiákból. Továbbá, mivel az alkalmazott módszerek statisztikai alapúak, nem felügyelt módszerekkel tanuló algoritmusokkal jönnek létre, ezért könnyen adaptálhatók tetszőleges doménre és nyelvre, ami a kézzel készített statikus erőforrásokról nem mondható el.

2. Kapcsolódó munkák

Számos módszer létezik szöveges korpuszokból hasonló szavak csoportjainak kinyerésére. A statisztikai módszerek általában a disztribúciós szemantika elméletét használják ki, azaz a szavak jelentését a környezetükben való előfordulásukhoz kötik. A különbség az egyes módszerek között leginkább a környezet definíciója. Egy lehetséges ábrázolás az egyes szavakhoz a függőségi relációk mentén kapcsolódó szavak használata környezetként, melyre több példát találunk a szakirodalomban (néhány ezek közül: [8], [7], [15] és [13]). Ezek alapja azonban egy jó minőségű függőségi elemző, vagy egy kézzel elemzett szöveges korpusz, amik közül gyakran egyik sem érhető el bizonyos nyelvekre. Más megvalósítások együttes előfordulások gyakorisági értékeiből hozzák létre a vektortérmodelleket és ezeket valamilyen vektorhasonlósági mérték alapján hasonlítják össze [16,3]. Napjaink egyre elterjedtebb módszere pedig a neurális hálózatok alkalmazása, amik egy kellően nagy szöveges korpuszból tanult folytonos vektorreprezentációt rendelnek az egyes szavakhoz, illetve a szavak jelentéséhez [10,11].

A jelen cikkben bemutatott megoldás abban különbözik a fentiektől, hogy a klaszterezés alapjául szolgáló, az egyes szavakat és kifejezéseket reprezentáló vektorok létrehozása során nem támaszkodtunk a szófaji egyértelműsítésnél mélyebb nyelvtani elemzésre, viszont nem is csupán a szavak együttes előfordulásainak statisztikáját vettük figyelembe.

3. Módszer

Az algoritmus három fő lépésből áll. Először a többszavas kifejezéseket összevonjuk a korpuszban. Ezeket aztán a későbbi lépések egyetlen egységként kezelik. A második lépés során létrehozzuk a disztribúciós modellt, azaz minden szópárhoz kiszámoljuk a disztribúciós hasonlóságuk mértékét. A harmadik lépésben ezeket a számértékeket használva jellemzőkként, minden kifejezéshez egy jellemzővektort hozunk létre, melyeket végül hierarchikus klaszterezéssel rendszerbe szervezünk. Ebből aztán tetszőleges sűrűséggel emelhetünk ki összetartozó kifejezéseket tartalmazó klasztereket.

3.1. Többszavas kifejezések

Mind az általános, mind a doménspecifikus nyelvhasználatban előfordulnak olyan kifejezések, amik több szóval írnak le egyetlen fogalmat. Függetlenül attól, hogy ezek szerkezete mennyire kompozicionális, önálló egységként kezelhetjük őket. Munkánk során a c-value algoritmus [6] módosított változatát használtuk a többszavas kifejezések azonosítására. Ennek bemeneteként szófaji egyértelműsítésen átesett szöveget használtunk, kimenetként pedig a többszavas kifejezések listáját kapjuk, a hozzájuk rendelt c-value értékek szerint sorba rendezve. Minél előrébb szerepel tehát egy kifejezés ebben a listában (azaz minél magasabb c-value értéket kapott), annál inkább tekinthető valódi többszavas kifejezésnek.

Az algoritmus menete a következő. Először kigyűjtjük a korpuszból az összes lehetséges n-grammot, ahol n értéke 1 és k közé esik (k tetszőlegesen választható, esetünkben $k = 20$). Ezután egy nyelvi szűrőt és egy stopword szűrőt alkalmaztunk, majd a szűrés után megmaradt n-grammokhoz meghatároztuk a korpuszbeli gyakoriságukat. Végül, a leghosszabbtól a legrövidebb n-grammok felé haladva kiszámítottuk a hozzájuk tartozó c-value értéket. Az algoritmus részletei megtalálhatók [17]-ben és [6]-ben.

A c-value érték meghatározása a korpuszból számított statisztikákra alapul, ezért nincs szükség külső lexikai erőforrások használatára a többszavas kifejezések megállapításához. Azonban a nyelvi szűrő kézzel definiált nyelvspecifikus szabályokat tartalmaz, annak biztosítására, hogy a kinyert kifejezések helyes kifejezések legyenek. Esetünkben ez a magyar főnévi szerkezetekre korlátozta a számításba jövő kifejezéseket, mivel kísérleteink során csak nominális fogalmakkal foglalkoztunk. Mivel a későbbi lépések során továbbra is szükségünk volt a szófaji címkékre, ezért az eredményként kapott összevont kifejezések a kifejezés fejének szófaji címkéjét kapták, megtartva ezáltal a szintaktikai szerepüket az adott kontextusban.

3.2. Disztribúciós hasonlóság

A releváns kifejezések csoportosításához szükség van egy hasonlósági metrikára is, ami két kifejezés jelentésbeli távolságát határozza meg. Erre szintén olyan nem felügyelt módszert alkalmaztunk, amely a hasonlóságokat nem egy külső erőforrás, ontológia alapján határozza meg, hanem a kifejezések korpuszbeli előfordulásai, az adott korpuszban való használatuk alapján.

A disztribúciós szemantika lényege, hogy a szemantikailag hasonló szavak hasonló környezetben fordulnak elő [5]. Tehát két szó jelentésének hasonlósága meghatározható a környezetük hasonlósága alapján. A szavak környezetét olyan jellemzőhalmazokkal reprezentáltuk, ahol minden jellemző egy relációból (r) és az adott reláció által meghatározott szóból (w') áll [8]. Ezek a relációk más alkalmazásokban általában függőségi relációk, mi azonban a függőségi elemző alkalmazásától most eltekintettünk. Carrol és tsai. [1] csupán a vizsgált szó meghatározott méretű környezetében előforduló szavak lexikai alakjának felhasználásával építettek ilyen szemantikai modellt. Mivel a mi esetünkben a morfológiai elemzés is rendelkezésre állt, ezért a következő jellemzőket vettük figyelembe:

- prev_1: a szót megelőző szó lemmája
- prev_w: a szó előtt 2-4 távolságon belül eső szavak lemmái
- next_1: a rákövetkező szó lemmája
- next_w: a szó után 2-4 távolságon belül eső szavak lemmái
- pos: a szó szófaja
- prev_pos: a szót megelőző szó szófaja
- next_pos: a szót követő szó szófaja

Szavak alatt pedig a lemmatizált szóalakot értjük a relációk mindkét oldalán.

Minden egyes jellemzőhöz meghatároztuk a korpuszbeli gyakoriságát. Ezekből a gyakoriságokból határozható meg a (w, r, w') hármas információtartalma $(I(w, r, w'))$ maximum likelihood becsléssel a következő képlettel:

$$I(w, r, w') = log \frac{||w, r, w'|| \times ||*, r, *||}{||w, r, *|| \times ||*, r, w'||}$$

Mivel $||w, r, w'||$ a (w, r, w') hármas korpuszbeli gyakoriságának felel meg, ezért ha a hármas bármelyik tagja $*$, akkor a hármas többi tagjára illeszkedő összes hármas gyakoriságának az összegével számolunk. Például a $||*, next\_1, ember||$ megfelel az olyan szavak gyakoriságának az összege, amit az *ember* szó követ.

Ezután a két szó (w_1 és w_2) közötti hasonlóságot a következő metrikával számoltuk [8] alapján:

$$SIM(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \bigcap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

ahol $T(w)$ azoknak az (r, w') pároknak a halmaza, ahol az $I(w, r, w')$ pozitív.

Bár a modell a korpuszban szereplő összes szóra alkalmazható, érdemes szófajonkénti modelleket építeni. Munkánk során csupán főnevekkel és nominális kifejezésekkel végeztünk kísérleteket, melyek legalább ötször előfordulnak a felhasznált korpuszban.

3.3. Hierarchikus klaszterezés

A szavak és kifejezések páronkénti hasonlóságából kiindulva fogalmi hierarchiát határozhatunk meg. Ehhez a leggyakoribb kifejezések és szavak csoportján agglomeratív klaszterezést hajtottunk végre. A klaszterező algoritmus megválasztásakor [14] érvelését vettük figyelembe, miszerint az írott szövegek kifinomult változatossága miatt a használt fogalmak csoportjainak száma előre nem megjósolható. Egy hierarchikus szerveződés azonban alkalmas arra, hogy az aktuális szövegre jellemző önálló, kompakt fogalmi csoportokat előre megfogalmazott általánosítás helyett az aktuális eredmény alapján nyerjük ki minden egyes szöveg esetén.

A legtöbb vektortéralapú módszer a fogalmak csoportosításakor azok együttes előfordulását veszi figyelembe, az egyes kifejezéseket leíró vektorok ilyen jellemzőket tartalmaznak. Ezek a megközelítések azonban alkalmatlanok arra, hogy olyan fogalmak között is felfedezzék a hasonlóságot, amik sosem fordulnak elő együtt, annak ellenére, hogy gyakran éppen az ilyen szavak azok, amik használatukat tekintve hasonlóak. Ezért az egyes kifejezéseket a többi kifejezéshez való hasonlóságukból álló jellemzővektorokkal ábrázoltuk. Így az egy kifejezéshez tartozó $c(w)$ vektor c_i eleme $SIM(w, w_i)$. Az egyes kifejezésekhez így létrehozott jellemzővektorokat klaszteresztük, ahol a klaszterek távolságát Ward ([18]) módszere alapján határoztuk meg. Ennek köszönhetően a kapott dendrogram alsó szintjein tömör, egymáshoz közel álló kifejezésekből álló csoportok jöttek létre.

Célunk azonban nem egy bináris faként ábrázolt teljes hierarchia meghatározása volt, hanem a fogalmak elkülönülő csoportjainak meghatározása, azaz a kapott dendrogram egyes kompakt részfái. Ezeket úgy kaphatjuk meg, ha a vágási pontokat a klaszterezés szintjei között lévő nagy ugrásoknál határozzuk meg. Formálisan ez úgy határozható meg, hogy a teljes fában lévő minden egyes részfát összekötő link magasságát összehasonlítjuk az alatta lévő szomszédos linkek magasságával egy adott mélységig. Ha ezek különbsége nagyobb, mint egy előre meghatározott küszöbérték (azaz a link inkonzisztens), akkor a vizsgált csomópont egy vágási pont. A teljes fából tehát az így meghatározott pontok alatti részfák levelei (azaz a szavak és kifejezések) egy csoportot alkotnak. Ezeknek a csoportoknak a sűrűsége a linkinkonzisztencia-küszöbérték változtatásával dinamikusan állítható.

4. Kísérletek

Kísérleteink során a Szeged Korpusz [2] 11 részkorpuszát használtuk a 1. táblázat szerint.

1. táblázat. A kísérletek során használt részkorpuszok, azok jellege és mérete

Név	**domén**	**Méret (token)**
10elb	diákfogalmazás	126841
8oelb	diákfogalmazás	92625
1984	szépirodalom	96843
utas	szépirodalom	75932
pfred	szépirodalom	60651
gazdtar	jog	153430
szerzj	jog	100153
newsml	rövid üzleti hírek	211742
mh	újság	49162
np	újság	74479
win2000	számítástechnika	66242

Minden egyes részkorpuszra egyesével alkalmaztuk a fenti algoritmust, azaz összevontuk az adott szövegtípusra jellemző többszavas kifejezéseket, ezután ezek hasonlóságát meghatároztuk, majd létrehoztuk a szemantikus csoportokat. Bár a Szeged Korpusz egyes részeinek mérete nem túl nagy, ezért az ezeken tanított statisztikai modellek kevésbé megbízhatóak, az összevont korpusszal is végeztünk kísérleteket, azonban ekkor sokkal kevésbé koherens csoportokat kaptunk eredményül.

Az alkalmazott algoritmus másik fő paramétere, ami hatással van az eredményként kapott klaszterek minőségére, a dendrogram vágási pontjait meghatározó inkonzisztenciaérték. Ezt az egyes részkorpuszoknál külön-külön állítottuk be a megfelelő eredmény elérése érdekében. Ennek a beállítása azonban függ az eredményül kapott csoportok felhasználási módjától, illetve a további feldolgozással kapcsolatos elvárásoktól. Ha nagyobb fedést szeretnénk elérni, akkor a küszöbérték magasabbra állítható, így nagyobb, de kevésbé tömör csoportokat kapunk. Ha azonban inkább kisebb, de szorosabban kapcsolódó kifejezéseket tartalmazó csoportokra van szükség, akkor a magasabb pontosság eléréséhez alacsonyabb küszöbértéket használhatunk. Mivel jelen munkánk során végzett kísérleteink csupán a módszer alkalmazhatóságát vizsgálták, a csoportok további feldolgozása nem volt meghatározva, ezért a küszöbérték beállítása empirikusan, a pontosság és a fedés közötti egyensúlyra törekedve történt minden egyes részkorpuszra.

Módszerünk eredményességét több síkon vizsgáltuk. Az egyik szempont a módszer domének közötti különbségekre való érzékenysége. Ehhez az eredményül kapott klasztereket az egyes részkorpuszokra páronként összehasonlítva egy kereszthasonlóság-értéket határoztunk meg minden párhoz. A részkorpuszok minden egyes lehetséges párosításához összegyűjtöttük azokat a szavakat és kifejezéseket, amik mindkét korpuszban előfordultak. Az ezeket a kifejezéseket tartalmazó csoportokra megvizsgáltuk a két részkorpusz esetén a csoportok metszetét és különbségét. A kapott halmazok méretének kumulált arányát a következő képlettel határoztuk meg:

$$\sum_{w} \frac{\|clust_A \bigcap clust_B\|}{\|clust_A\| + \|clust_B\| - \|clust_A \bigcap clust_B\|}$$

ahol $w \in (text_A \bigcap text_B)$ és $clust_A$ és $clust_B$ a két vizsgált részkorpusz klaszterei, amikben a w kifejezés előfordul. A 2. táblázat tartalmazza az így kapott kereszthasonlóság érték szerint rendezett lista első és utolsó 5 elemét.

Ahogy az eredményekből látszik, az egymáshoz közelebb álló domének szerepelnek a lista elején, míg az egymástól távolabbi párok kerültek a rangsor végére. A párok közötti eltérés nem csak a bennük előforduló különböző szavakból fakad (a `mh` és a `newsml` pár metszetében szereplő szavak száma közel azonos a `10elb` és a `gazdtar` pár metszetében szereplő szavak számával, mégis az előbbi pár a lista elején szerepel, míg az utóbbi a végén), hanem az azonos szavak különböző használatából is. Ezekre a különbségekre nem derülhet fény, egy előre definiált erőforrás alapján, hiszen abban nem különböztetjük meg a különböző típusú szövegekben megjelenő kisebb jelentésbeli vagy hangsúlybeli különbségeket.

2. táblázat. A kereszthasonlóság (KH) vizsgálatának eredményei

Pár	**KH**	**domén**
10elb–8oelb	10.880	diákfogalmazás–diákfogalmazás
newsml–np	4.586	hír–újság
mh–newsml	4.500	újság–hír
mh–np	3.672	újság–újság
gazdtar–szerzj	3.513	jog–jog
10elb–np	2.223	diákfogalmazás–újság
1984–pfred	2.179	szépirodalom–szépirodalom
8oelb–utas	2.057	diákfogalmazás–szépirodalom
10elb–newsml	1.917	diákfogalmazás–hír
...	...	...
8oelb–szerzj	0.321	diákfogalmazás–jog
mh–pfred	0.321	hír–szépirodalom
pfred–szerzj	0.222	szépirodalom–jog
gazdtar–utas	0.192	jog–szépirodalom
10elb–gazdtar	0.182	diákfogalmazás–jog
pfred–win2000	0.154	szépirodalom–számítástechnika
8oelb–win2000	0.000	diákfogalmazás–számítástechnika
gazdtar–pfred	0.000	jog–szépirodalom
utas–win2000	0.000	szépirodalom–számítástechnika

5. Eredmények

Az eredményül kapott fogalmi klaszterek különböző szempont szerinti csoportosításokat eredményeztek. Néhány csoportban általános vagy klasszikus értelemben kapcsolódó kifejezések gyűltek össze, mint például testrészek (ezek elsősorban szépirodalmi szövegekben jelentek meg, ahol az egyes szereplők leírása részletesebb), napok és hónapok nevei (elsősorban a hír és a diákfogalmazás részkorpuszokban) vagy pénznemek (a gazdasági és hírkorpuszokban). Bár ezek az általános csoportok akár előre is meghatározhatók, nincs garancia arra, hogy nem jelenik meg egy olyan kifejezés egy adott szövegben, ami eredetileg nem lenne benne az előre definiált listákban, így ezeket is érdemesebb az adott szövegből kinyerni. Továbbá a kinyert csoportok nem tartalmaznak olyan szavakat és kifejezéseket, amik az adott szövegben nem szerepelnek, így az eltárolandó eredmény mérete sem haladja meg azt, amire feltétlenül szükség van.

A létrejött csoportok egy másik típusa valamilyen nyelvtani szempont szerinti rendeződés alapján jött létre, mint például a funkcióigés szerkezetek főnévi magját alkotó elemek.

A harmadik fő típusba pedig olyan csoportok sorolhatók, amikben a szavak valamilyen tágabb értelemben kapcsolódnak, leginkább az adott részkorpuszra jellemző használatuk alapján. Néhány ilyen példát láthatunk a 3. táblázatban.

Ahogy a példákon is látszik, az alkalmazott algoritmus sokszor valamilyen asszociációs kapcsolatban álló kifejezéseket csoportosított össze, különösen a diákfogalmazás és a szépirodalmi részkorpuszok esetén. Például a *erdő, falu, város,*

3. táblázat. Néhány példa az eredményül kapott csoportokra az egyes részkorpuszokból

Text	**cluster**
gazdtar	*vezető tisztségviselő, könyvvizsgáló, személy, igazgatóság, ügyvezető, igazgató*
gazdtar	*társasági szerződés, alapító okirat, alapszabály*
10elb	*erdő, falu, város, ház, diszkó, part*
10elb	*cucc, táska, csomag, holmi*
1984	*ujj, test, arc, szem, fej, kar, kéz, tömeg, agy, száj, láb*
1984	*férfi, asszony, pillanat, hang, telekép, lány, ember, pont, Mr., éves kor*
1984	*lázadás, szokás, remény, napló, hit, dátum*
8oelb	*öröm, élmény, irány, nyaralás, történet, délután*
newsml	*költség, kiadás, díj, adósság, befektetés, eszköz*
newsml	*fél, egész, arány, időszak*
szerzj win	*fejezet, cikk, pont, törvény, §, bekezdés* *NTFS, állományrendszer, helyfoglalási egység, adat, lemez, logikai lemez, kötet, merevlemez, fizikai lemez*

ház, diszkó, part csoportban a kifejezések páronkénti kapcsolata nem feltétlenül megjósolható (pl. az *erdő* és *diszkó* pár esetén), de ismerve a részkorpuszt (diákok által írt szövegek), illetve a csoportba sorolt többi szót, már könnyen belátható, hogy a csoportosításnak van értelme, a kifejezések valóban kapcsolódnak egymáshoz. Egy másik jellemzője az alkalmazott algoritmusnak, hogy könnyen alkalmazkodik a doménspecifikus, vagy akár teljesen egyedi szóhasználathoz is. Például a diákfogalmazásokra jellemző szleng is megfelelően csoportosítható. Ezeket a szóalakokat szinte lehetetlen egy előre definiált kategóriarendszerbe besorolni, hiszen nagyon gyorsan jelennek meg, vagy tűnnek el a nyelvből, esetleg átalakul a jelentésük. Egy másik példa a szépirodalmi szövegekből alkotott csoportosítások esetén látható, különösen George Orwell *1984* című regénye esetén. Ez a korpusz rengeteg sajátos szóalakot tartalmaz, amik csupán a szerző által kitalált, a valóságban nem létező, vagy nem az ebben a műben használt értelemben használt kifejezések, az alkalmazott algoritmus azonban ezeket is helyesen tudta csoportosítani, a ténylegesen létező szavakkal együtt az általánostól esetlegesen eltérő, éppen megfelelő jelentésük szerint (pl. *lázadás, szokás, remény, napló, hit, dátum*).

Az eredmények vizsgálata azonban nem csak az egyes részkorpuszok esetén érdekes, hanem a létrejött csoportosítások metszetét és különbségeit is érdemes elemezni. Például az *autó* szó több részkorpuszban is a családtagokat leíró csoportba került besorolásra. Szigorúan szemantikai szempontból ennek a relációnak

nincs értelme, ugyanakkor a valóságban gyakran tényleg létező jelenség az autóra mint családtagra való utalás. A diákfogalmazások esetén pedig még a *bicikli* szó is ebbe a csoportba került, ami hasonlóan magyarázható. Megfigyelhetőek továbbá a különböző domének közötti apró eltolódások is a szóhasználatot illetően. Például a 8. osztályos diákok által írt fogalmazásokban a *szülő* és a *barát* szavak még egy csoportba kerültek, azonban a tizedikes diákok által írt fogalmazásokban ez a két szó már elválik, ami jól tükrözi a gyerek-szülő viszony eltolódását ennél a korosztálynál.

6. Konklúzió

Jelen cikkünkben olyan kísérletekről számoltunk be, amelyek azonos fogalmak különböző jellegű szövegekben való használatát vizsgálják. Ehhez eszközül a disztribúciós szemantika egy modelljét alkalmaztuk. A többszavas kifejezések meghatározása után minden szót/kifejezést a többi szóhoz való hasonlóságát tartalmazó vektorral ábrázoltunk (ahol a páronkénti hasonlóság számítása a kölcsönös információtartalmon alapult [8]). Az így kapott vektorokat pedig hierarchikus klaszterezéssel tömör, koherens csoportokba osztályoztuk. Az eredményül kapott csoportok tehát olyan kifejezéseket és szavakat tartalmaznak, amelyek használatuk szempontjából hasonlóak.

A fenti algoritmust a Szeged Korpusz [2] egyes részkorpuszaira külön-külön alkalmaztuk. Az eredmények elemzésekor pedig azt vizsgáltuk, hogy ugyanazon kifejezések disztribúciós viselkedése hogyan változik különböző domének esetén. Így olyan kifinomult különbségekre is fény derült, melyek semmilyen formális ontológiában vagy fogalmi rendszerben nem ábrázolhatóak.

A módszerünk ellenőrzéseként definiáltunk egy olyan metrikát, ami a különböző doménekből létrejött csoportok közötti átfedés mértékét vizsgálja. Ezzel kimutattuk, hogy a hasonló jellegű (gazdasági-jogi, sajtónyelvi, szépirodalmi, iskolai) szövegekből épített fogalmi csoportok nagyobb átfedést mutattak, mint a különböző domének fogalmi csoportjai.

Hivatkozások

1. Carroll, J., Koeling, R., Puri, S.: Lexical acquisition for clinical text mining using distributional similarity. In: Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II. pp. 232–246. CICLing'12, Springer-Verlag, Berlin, Heidelberg (2012)
2. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Sojka, P., Kopecek, I., Pala, K. (eds.) TSD. Lecture Notes in Computer Science, vol. 3206, pp. 41–48. Springer (2004)
3. de Cruys, T..: Semantic clustering in Dutch. In: Proceedings 16th Meeting of Computational Linguistics in the Netherlands. pp. 19–31 (2005)
4. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
5. Firth, J.R.: A Synopsis of Linguistic Theory, 1930-1955. Studies in Linguistic Analysis pp. 1–32 (1957)

6. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries 3(2), 115–130 (August 2000)
7. Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the 28th Annual Meeting on Association for Computational Linguistics. pp. 268–275. ACL '90, Association for Computational Linguistics, Stroudsburg, PA, USA (1990), http://dx.doi.org/10.3115/981823.981857
8. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics - Volume 2. pp. 768–774. COLING '98, Association for Computational Linguistics, Stroudsburg, PA, USA (1998)
9. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of The Fourth Global WordNet Conference. pp. 311–321 (2008)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
11. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (June 2013), http://www.aclweb.org/anthology/N13-1090
12. Miller, G.A.: WordNet: A Lexical Database for English. COMMUNICATIONS OF THE ACM 38, 39–41 (1995)
13. Padó, S., Lapata, M.: Dependency-based construction of semantic space models. Comput. Linguist. 33(2), 161–199 (Jun 2007), http://dx.doi.org/10.1162/coli.2007.33.2.161
14. Pereira, F., Tishby, N., Lee, L.: Distributional Clustering of English Words. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics. pp. 183–190. ACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA (1993), http://dx.doi.org/10.3115/981574.981598
15. Ruge, G.: Experiment on linguistically-based term associations. Inf. Process. Manage. 28(3), 317–332 (Jan 1992), http://dx.doi.org/10.1016/0306-4573(92)90078-E
16. Senellart, P., Blondel, V.: Automatic discovery of similar words. In: Berry, M. (ed.) Survey of Text Mining. Springer-Verlag (2003)
17. Siklósi, B., Novák, A.: Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods, Lecture Notes in Artificial Intelligence, vol. 8791, pp. 233–243. Springer International Publishing, Heidelberg (2014)
18. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963), http://www.jstor.org/stable/2282967
19. Zhang, J.: Representations of health concepts: a cognitive perspective. Journal of Biomedical Informatics 35(1), 17 – 24 (2002), http://www.sciencedirect.com/science/article/pii/S1532046402000035